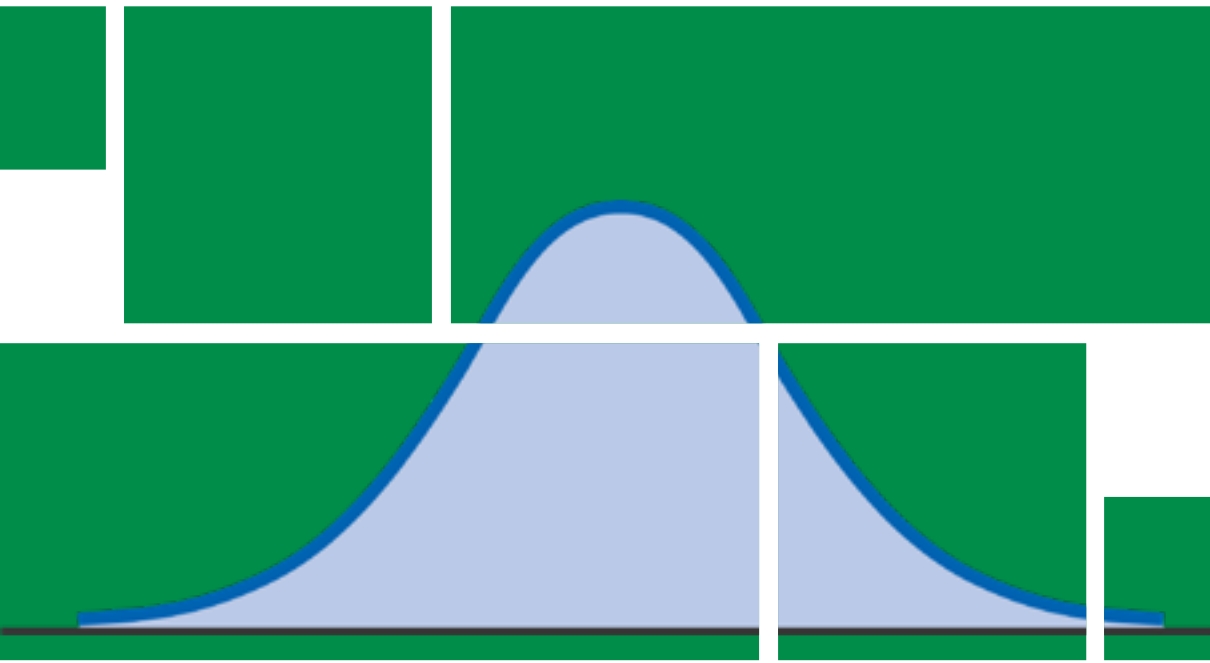
ISSN 1809-4996
Maio / 2019

DOCUMENTOS

233

Ensaios:

Da amostra ao teorema do limite central
Um pouco dos fundamentos e uma aplicação prática



Embrapa

ISSN 1809-4996
Maio/2019

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Mandioca e Fruticultura
Ministério da Agricultura, Pecuária e Abastecimento

DOCUMENTOS 233

Ensaios:

Da amostra ao teorema do limite central
Um pouco dos fundamentos e uma aplicação prática

Clóvis Oliveira de Almeida
(Autor)

Embrapa Mandioca e Fruticultura
Cruz das Almas, BA
2019

Exemplares desta publicação podem ser adquiridos na:

Embrapa Mandioca e Fruticultura
Rua Embrapa, s/nº, Caixa Postal 07
44380-000, Cruz das Almas, Bahia
Fone: 75 3312-8048
Fax: 75 3312-8097
www.embrapa.br
www.embrapa.br/fale-conosco/sac

Supervisão editorial
Francisco Ferraz Laranjeira

Revisão de texto
Adriana Villar Tullio Marinho

Normalização bibliográfica
Lucidalva Ribeiro Gonçalves Pinheiro

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Anapaula Rosário Lopes

1ª edição
On-line (2019).

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Mandioca e Fruticultura

Almeida, Clóvis Oliveira de
Ensaios: Da amostra ao teorema do limite central Um pouco dos fundamentos e uma aplicação prática / Clóvis Oliveira de Almeida – Cruz das Almas, BA : Embrapa Mandioca e Fruticultura, 2019.
40 p. il. ; 21 cm. - (Documentos/ Embrapa Mandioca e Fruticultura, ISSN 1809-4996.233).

1. Estatística. I. Título. II. Série.

CDD 519.5

Apresentação

Nesta publicação, o autor aborda, de forma clara e concisa, duas das mais importantes descobertas da estatística: o processo de amostragem aleatória e o teorema do limite central, bem como a conexão entre eles. Tudo isso é feito sem recorrer a equações matemáticas, porque o propósito maior é justamente tornar o texto amigável e compreensível a um público que extrapole os limites da academia. Para tanto, o autor tenta evitar ao máximo o uso de termos técnicos, mas, quando o mesmo se torna inevitável, ele é imediatamente seguido de uma definição básica ou de um exemplo no mundo real. O leitor vai "viajar" em uma breve história relacionada a tais descobertas e talvez se surpreenda ao saber que muitas mentes brilhantes de matemáticos e físicos estiveram envolvidas em seu desenvolvimento por aproximadamente dois séculos. Não é à toa que o teorema do limite central desfruta de grande poder teórico e de aplicação prática em várias áreas do conhecimento. O autor finaliza o documento com um exemplo de aplicação prática a uma situação real de grande interesse da Embrapa, dos agentes ligados à agricultura brasileira e dos formuladores de políticas públicas relacionadas ao setor.

Alberto Duarte Vilarinhos

Chefe-geral da Embrapa Mandioca e Fruticultura

Sumário

Resumo

Este documento resume em linguagem simples e direta para o grande público não especializado em estatística e disciplinas correlatas, as ideias básicas de duas das mais grandiosas descobertas da Estatística: o processo de amostragem aleatória e o teorema do limite central. Após apresentar as ideias básicas que fundamentaram ambas as descobertas, ao final do Documento é fornecido um exemplo de aplicação prática entre várias outras possibilidades de utilização do teorema do limite central, que tem sido, por mais de meio século, uma das principais bases das ferramentas de análise de dados em diversas áreas do conhecimento, das ciências físicas às ciências humanas. O exemplo apresentado deixa evidente os riscos de se fazer inferências sobre a população a partir de uma amostra não aleatória: a perda da acurácia. A mensagem básica por trás do processo de amostragem aleatória é que cada membro de uma população tenha as mesmas chances de ser selecionado. Essa condição é a melhor forma de se evitarem os vieses no processo de amostragem em pesquisas básicas e aplicadas. O teorema do limite central, por sua vez, nos diz que uma amostra grande o suficiente, escolhida de maneira aleatória, na grande maioria das vezes, representará bem a população da qual foi extraída. Além disso, esse mesmo teorema nos assegura que as médias dessas mesmas amostras têm distribuição aproximadamente normal em torno da média da população, independentemente do tipo de distribuição de probabilidade subjacente aos dados. E é justamente daí que emergem toda simetria, poder, magia e beleza do teorema do limite central.

Palavras-chaves: randomização, probabilidade, grandes amostras, inferência, censo agropecuário.

Introdução

À primeira vista, parece estranho que um documento publicado pela Embrapa Mandioca e Fruticultura trate de uma questão puramente estatística. Mas essa impressão somente se manifesta à primeira vista e, de certa forma, ela é natural porque, no nosso dia a dia, o produto de nosso trabalho deve ser o mais concreto e o mais visível possível aos olhos de nosso público-alvo: os produtores rurais e os setores a eles conectados. Após entender o poder destas duas grandiosas descobertas estatísticas, a amostragem aleatória e o teorema do limite central, e perceber que, de certa forma, elas estão presentes na parte não visível de quase todos os nossos produtos físicos, o leitor acaba se convencendo da importância dessas duas descobertas nas pesquisas básicas e aplicadas, de uma forma geral, e nas pesquisas agronômicas, de uma forma específica. O processo de amostragem aleatória é essencial em pesquisas para evitar os vieses de seleção que poderiam levar a resultados enganosos e invalidar os resultados da própria pesquisa. Portanto, as boas pesquisas que têm por base a amostragem precisam, sempre que possível, ser fundamentadas em amostras suficientemente grandes e aleatórias, nas quais cada elemento de uma dada população tenha as mesmas chances de ser selecionado. O teorema do limite central permite, com base em amostras assim selecionadas, ir ainda muito mais além: ele nos assegura que as médias das amostras assim selecionadas apresentem distribuição aproximadamente normal em torno da média da população, independentemente do tipo de distribuição subjacente aos dados. E isso faz toda a diferença em análise matemática e estatística: a distribuição normal torna a estatística mais facilmente aplicável ao mundo real. Muitas vezes, mesmo sem nos darmos conta disso, utilizamos do poder do teorema do limite central rotineiramente em quase todas as nossas atividades de pesquisa. Ele está presente nos testes de diferença entre médias, intervalos de confiança e em várias formas de aplicação no mundo corporativo disseminadas nos conceitos apresentados por Walter Shewhart[1] e sintetizados na forma de controle estatístico da qualidade.

[1] Ver Deming (2003); Rodrigues (2006) e Montgomery (2017), para uma vasta abordagem e exemplos de aplicação do controle estatístico da qualidade.

A minha amostra é verdadeiramente aleatória[2]?

O título deste item remete a uma questão que sempre deveria ser feita e respondida por um pesquisador ao planejar e ao analisar os dados de uma pesquisa ou de qualquer coleção de dados que envolva a amostragem e a análise estatística. A amostra aleatória é uma condição necessária na elaboração e na validação de estudos científicos. Embora pareça óbvio, uma grande parte de nossos pares não seleciona adequadamente o que se poderia chamar de uma verdadeira amostra[3] aleatória, seja em virtude do desconhecimento do real significado do termo, seja por questões de custos, tempo ou dificuldades operacionais envolvidas no processo de seleção de uma amostra aleatória.

Em geral, sobretudo em pesquisas de opinião, dá-se maior importância ao tamanho da amostra que à sua forma de seleção, com base na crença ilusória de que amostras maiores sempre seriam representativas de uma dada população. Mas, infelizmente, nem sempre esse é o caso. Por definição, uma amostra aleatória ou randomizada, como também é conhecida, é aquela em que cada elemento de uma população tem as mesmas chances de ser selecionado, ou seja, em uma amostra aleatória não pode existir nenhum tipo de favorecimento ou restrição, ela deve ser totalmente imparcial. É como se deixássemos o acaso operar a nosso favor. Uma vez ou outra ele também pode “falhar” e produzir coisas que não se “pareçam” nada com o acaso, embora ainda assim o seja. Aliás, se a teoria da evolução de Charles Darwin estiver correta, todos nós, de certa maneira, também somos frutos do acaso: do casamento entre as mutações aleatórias (os erros de cópia, a face casual) com a seleção natural (a face não casual). Lembre-se de que coisas improváveis também acontecem, embora sejam extremamente raras. O filósofo grego Aristóteles, que viveu antes da Era Cristã, já tinha percebido que “é da natureza da probabilidade que coisas improváveis

[2] O texto que segue trata do método de amostragem aleatória simples, que é apenas um dos principais procedimentos de amostragem probabilística: aleatória sistemática, estratificada e por conglomerados.

[3] Em estatística, uma amostra pode ser definida como um conjunto formado por um subconjunto da população. A população, por sua vez, é o conjunto formado por todos os elementos que possuem pelo menos uma característica em comum, como por exemplo: população de eleitores de um país, população de municípios produtores de laranja de um determinado estado, população brasileira com renda acima de três salários mínimos etc.

aconteçam”. Mas, na média, em estatística, ele (o acaso) é implacável e vai terminar por produzir um resultado tal que tende a reproduzir em amostras suficientemente grandes e aleatórias as características da população que elas representam. Parece até ironia, mas a escolha ao acaso de uma amostra é a forma mais racional de selecionar os indivíduos (ou elementos) para representar a população à qual eles pertencem. Isso acontece porque somente a escolha ao acaso é capaz de proporcionar a cada indivíduo da população as mesmas chances de ser selecionado. Portanto, para que uma amostra seja boa, não basta que ela seja grande, ela precisa ser escolhida de forma aleatória: uma amostra maior não é capaz de compensar os vieses causados por erros de seleção.

Na área médica, por questão de dificuldades operacionais ou éticas, muitos são os exemplos de pesquisas que não se baseiam em amostras aleatórias. Não se pode simplesmente sair por aí selecionando pessoas a partir de uma lista para participar de uma pesquisa que tenha como objetivo, por exemplo, testar a eficiência de uma droga qualquer sobre o que quer que seja. Portanto, na área médica, em geral, os estudos são baseados em voluntários que se candidatam para participar de uma pesquisa. Nesses casos, o melhor a ser feito é selecionar os voluntários com base nas características da população alvo que se deseja representar, embora se faça uso da aleatoriedade para selecionar as clínicas ou os hospitais onde os pacientes são tratados.

De modo geral, os principais exemplos de amostras não aleatórias estão associados às enquetes feitas por internet, televisão, telefone, rádio e, também, pelas pesquisas com voluntários, simplesmente porque, em todos esses casos, por diferentes motivos, não são dadas a cada um dos membros da população alvo as mesmas chances de participar da pesquisa. Mas, felizmente, esse não é o caso das pesquisas agronômicas, biológicas, físicas e de muitas outras pesquisas de opinião.

Por esse motivo, voltaremos nossa atenção aos casos em que a pesquisa pode ser feita com base em amostra aleatória, embora muitas vezes também esses casos sejam negligenciados. Uma das principais fontes de vieses em pesquisas é o uso de amostras selecionadas de maneira não aleatória. Não importa o tamanho que sua amostra tenha, se ela não for selecionada aleatoriamente, os vieses estarão presentes nos dados, e a presença de

vieses em estudos científicos invalidam os resultados da pesquisa, uma vez que não se pode fazer generalizações e, portanto, extrair resultados conclusivos. Como diria a professora Deborah Rumsey[4], diretora do Centro de Aprendizagem de Matemática e Estatística da Universidade do Estado de Ohio, nos Estados Unidos: se a amostra não for selecionada de maneira aleatória, os resultados equivalem a um grão de areia. Assim, chegamos a outro extraordinário poder da estatística, que está associado à possibilidade de extrapolação dos resultados de um estudo científico, também conhecido como inferência. Uma boa parte desse poder, quase mágico, vem da aleatorização ou da randomização no processo de seleção da amostra; a outra, do próprio tamanho da amostra, e tudo isso acaba convergindo para a "magia" e a beleza proporcionadas pelo teorema do limite central, que, em essência, nos diz que uma amostra grande o bastante, escolhida de maneira adequada, na grande maioria das vezes, representará bem a população da qual foi extraída.

A forma mais fácil de coletar uma amostra representativa de uma população é selecionar de maneira aleatória um subconjunto dessa população. Com base no princípio da aleatorização e do controle, o ainda jovem estatístico Ronald Fisher[5], considerado o fundador da estatística moderna, fez uma verdadeira revolução científica quando trabalhava na Estação Agrícola Experimental de Rothamsted, hoje rebatizada com o nome Rothamsted Research, um centro de pesquisa localizado na cidade de Harpenden, Hertfordshire, na Inglaterra, o berço da estatística experimental e da agricultura moderna. Aliás, devemos a Fisher o que hoje se conhece por estatística experimental. Ao fazer uso do princípio da randomização e do controle, Fisher resolveu um problema secular dos mais desafiadores das pesquisas agronômicas da então Estação Agrícola Experimental de Rothamsted. Antes dele, um único "experimento" para testes de resposta de variedades à fertilização era conduzido em grandes áreas e em anos a fio, sem o uso de testemunhas ou grupo de controle, o que levou à perda de tempo e de recursos financeiros porque não se tinha como separar o efeito da fertilização dos demais efeitos sobre as variedades testadas ou sobre o grupo de tratamento. O resultado foi uma perda quase secular de "experimentos". Ele simplesmente descobriu

4 Rumsey (2016)

5 Ver Salsburg (2009), para maiores detalhes acerca deste importante registro histórico.

- digo simplesmente agora quando é possível uma análise retrospectiva, uma vez que sabemos o que foi feito - que experimentos cuidadosamente randomizados e controlados dispensariam o uso de grandes áreas e permitiria a comparação de diferentes tratamentos em um mesmo ano. As grandes áreas foram substituídas por parcelas, cada uma contendo fileiras de plantas e cada fileira recebendo um tratamento diferente. Uma das grandes sacadas de Fisher foi perceber que a aleatoriedade poderia resolver para nós os problemas decorridos das diferenças de gradiente de fertilidade e de outras propriedades do solo, equilibrando-as para nós entre as diferentes variedades em pequenas fileiras experimentais. Ou seja, em essência, esse é o papel da aleatoriedade: equilibrar as diferenças relevantes (visíveis e não visíveis) entre os grupos ou indivíduos de uma amostra. O uso do controle (a testemunha, que não recebeu o tratamento) desempenhava o papel que permitiria atribuir a causa das variações ou efeitos. Os princípios estabelecidos por Fisher àquela época norteiam até hoje a estatística experimental: a aleatorização, o controle local e a repetição. Para contornar o efeito do acaso sobre os resultados do experimento, um papel também desempenhado pela randomização, a repetição utiliza o efeito médio obtido em cada tratamento.

Embora este tema esteja longe de ser esgotado, gostaríamos de finalizá-lo deixando uma dica que talvez seja útil para o leitor interessado que não tenha conhecimento suficiente na área, é claro, qual seja: para muitas de nossas pesquisas aplicadas, uma forma rápida e prática de fazer uma seleção amostral de maneira aleatória (seja de plantas, sementes, folhas, frutos, raízes, animais, indivíduos ou mesmo localidades etc.) seria criar uma lista associando cada elemento observacional a um número. Feito isso, bastaria utilizar depois um programa estatístico para fazer a escolha aleatória, sem repetição, dos elementos da amostra. Tais procedimentos de aleatorização são também facilmente realizados e encontrados em aplicativos de celulares.

Enquanto geradores de conhecimentos e informações que podem e devem ser utilizados para transformar vidas, convido-os a uma reflexão sobre nossas práticas cotidianas quanto ao rigor necessário na seleção de amostras. Talvez resida justamente aí o porquê de nem sempre os resultados das pesquisas alcançarem os objetivos esperados na vida real.

Teorema do Limite Central

O teorema do limite central, ou teorema central do limite, como também é conhecido, é nada mais do que um dos mais importantes e marcantes teorema da Estatística. Aliás, o termo "teorema do limite central" dá uma ideia do quão importante ele é. Muitos foram os matemáticos que contribuíram para o surgimento deste belo e importante teorema. Ele não é obra de uma única mente proeminente, tampouco de uma época: foram mais de 197 anos desde a sua primeira concepção até o seu pleno desenvolvimento[6]. A primeira formulação do teorema devemos ao matemático francês Abraham de Moivre, datada de 1733. Depois, foi a vez do genial matemático alemão Carl Friedrich Gauss, conhecido como o "príncipe da matemática", que até hoje tem o seu nome ligado à curva normal, ou gaussiana, na qual o teorema se fundamenta. Em seguida, surgiram as contribuições de outro famoso matemático francês, Pierre Simon de Laplace, em um trabalho publicado em 1812, que teve como fonte de inspiração o trabalho de Gauss. Em 1901, o matemático russo Aleksandr Lyapunov provou como o teorema do limite central funcionava matematicamente. Mas, somente no começo dos anos 1930, o matemático francês Paul Lévy determinou as condições gerais de funcionalidade do teorema[7]. Porém, muito antes de todos eles, outros matemáticos deram importantes contribuições para a construção desse teorema, destacando-se entre eles o francês Blaise Pascal, que antecedeu Moivre, e o suíço Damiel Bernoulli, que precedeu os demais. Devemos a esse teorema muito do poder dessa que se poderia chamar de "irmã" mais jovem da Matemática. Antes de se tornar a "irmã" da Matemática, a Estatística tinha o seu nome associado a pessoas ligadas a jogos de azar e, talvez, por isso mesmo, tenha demorado tanto tempo para ser reconhecida com um importante ramo da Matemática: a Estatística Matemática.

Em essência, o teorema do limite central nos diz que as médias de amostras suficientemente grandes, selecionadas de maneira adequada, apresentam distribuição aproximadamente normal em torno da média da população, independentemente do tipo de distribuição de probabilidade subjacente aos dados[8]. É justamente a passagem proporcionada, em grande medida, pelas

[6] Ver MLODINOW, 2009; SALSBURG, 2009; ELLENBERG, 2015, para maior riqueza de detalhes históricos acerca do teorema do limite central.

[7] Essas condições podem ser resumidas em apenas duas, e são adequadas em situações nas quais têm-se uma sequência de números gerados aleatoriamente, um após o outro; quais sejam: variância finita e passeio aleatório. A essa sequência Lévy deu o nome de acumulada (ver Salsburg, 2009, p. 221).

[8] O tamanho de cada amostra para que o teorema do limite central se verifique depende do tipo de distribuição de probabilidade subjacente aos dados. Com base na regra prática, se o tamanho de cada amostra for maior do que 30, a distribuição das médias amostrais pode ser aproximada satisfatoriamente por uma distribuição normal.

médias de qualquer tipo de distribuição para a normal, onde reside toda a simetria, poder, magia e beleza do teorema do limite central. A distribuição normal é o alicerce desse teorema. Poder desfrutar da normalidade no padrão de distribuição de dados faz toda a diferença em análise estatística, "porque assim a matemática é tratável", já dizia o professor e estatístico David Salsburg[9]. Os dados que seguem a distribuição normal, também conhecida como gaussiana, ou curva em forma de sino, estão distribuídos de forma simétrica ao redor de uma média, que é representada pelo pico da curva; aproximadamente 68% das observações estarão dentro do limite de 1 desvio padrão da média; cerca de 95%, dentro de 2 desvios padrão; e 99%, dentro de 3 desvios padrão (Figura 1). Ou, dito de outra maneira, à medida que nos afastamos da média (ou do centro da curva), menos dados serão encontrados em ambos os lados (Figura 1). Em uma distribuição normal, os dados que ocorrem com menor frequência são os de valores atípicos e eles estão distribuídos de forma simétrica a mais de 2 desvios padrão da média, nos dois extremos ou caudas da curva, e, em virtude disso, a média populacional não será influenciada por esses valores atípicos (Figura 1). O desvio padrão é uma medida de variabilidade, ou dispersão, expressa em unidade de medida do dado original, e pode ser calculado, simplesmente, extraindo-se a raiz quadrada de outra medida de variabilidade: a variância. O desvio padrão nos fornece a distância média do conjunto de dados em relação à média, ou ao centro da distribuição (Figura 1).

Então, com base no teorema do limite central, podemos afirmar que a melhor hipótese acerca do valor esperado da média de qualquer amostra de tamanho suficientemente grande e escolhida de maneira aleatória é a média da população à qual ela pertence. Daí emergem as diversas possibilidades de utilização prática do teorema do limite central. Isso porque, uma vez conhecida a média populacional ou, equivalentemente como nos ensina o próprio teorema, a média das médias amostrais, pode-se fazer uso desse teorema, por exemplo, para verificar a consistência entre a média populacional e uma média amostral qualquer, supostamente extraída (a amostra) de forma adequada da população alvo. Ou seja, o teorema do limite central permite calcular a probabilidade de que uma média amostral específica tenha sido extraída de uma dada população. Se essa probabilidade for baixa, conclui-se que, muito provavelmente, a referida média não tenha sido extraída da população alvo, ou que a mesma não poderia representar a média populacional.

[9] Salsburg (2009, p. 81-82).

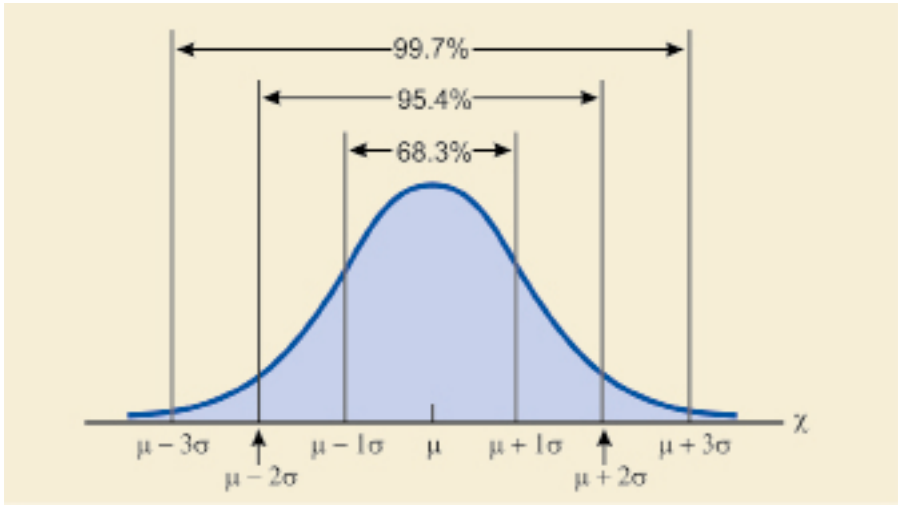


Figura 1. Curva de distribuição normal, curva de Gauss ou curva em forma de sino de uma sequência de dados
Fonte: ESCOLAEDIT (2018)

Mas como é possível substituir a média populacional pela média das médias amostrais? E, ainda, como é possível utilizar desse princípio para se fazer perícia em dados? O teorema do limite central nos assegura que, se as amostras forem suficientemente grandes e escolhidas de maneira aleatória, as médias de todas as amostras possíveis estarão distribuídas simetricamente em torno da média populacional. Assim sendo, a Matemática nos assegura que a média das médias das possíveis amostras terá um valor igual à média populacional. Embora isso nos dê algum alento, à primeira vista parece, mas apenas parece, ter pouca aplicabilidade, seja por questões de custos ou de tempo para se obterem todas as possíveis amostras necessárias ao cálculo da média das médias amostrais.

A lei dos grandes números, outro importante teorema da Estatística, publicado em 1713, após a morte de seu criador Jakob Bernoulli, também membro da talentosa e notória família Bernoulli, de matemáticos, nos ensina que não é preciso conhecer todas as possíveis médias amostrais, porque, à medida que aumenta o número de tentativas (no caso, as amostras colhidas), a média dos resultados vai se aproximando cada vez mais do valor esperado (a média populacional). Ou, dito de outra forma, a probabilidade de se obter

uma média das médias que se desvie excessivamente da média populacional decresce rapidamente à medida que o número de tentativas aumenta (o número de amostras). A probabilidade está no “DNA” da Estatística para lidar com a incerteza. Ela funciona muito bem para prever comportamento de longo prazo ou com várias tentativas; mas usualmente “falha” a curto prazo, ou quando só é possível se fazer poucas ou apenas uma tentativa.

O fenômeno da regressão à média, uma descoberta de 1875, do matemático Inglês Francis Galton, é outra evidência disso. Ele descobriu que em medidas relacionadas de qualquer série de eventos aleatórios, há uma grande probabilidade de um acontecimento extremo ser seguido, em virtude puramente do acaso, por um acontecimento mais próximo da média, ou que um acontecimento mais próximo da média preceda, também por puro acaso, um acontecimento extremo. Eis aí a origem de uma das grandes fragilidades dos estudos fundamentados na metodologia do antes e do depois, sem o uso devido do controle ou da testemunha. A regressão à média também é uma explicação adicional para justificar, na área médica, a aferição da pressão arterial a partir de várias medições, sendo três o número mínimo recomendado. Se uma medição inicial apresentar valores extremos, é mais provável que as medições subsequentes se aproximem da média, fornecendo uma leitura mais representativa da pressão arterial geral. Portanto, é necessário que se colete um maior número de observações e amostras para se ter uma ideia melhor da distribuição de probabilidade dos eventos aleatórios. Ou seja, a probabilidade opera no curso de seu próprio tempo para prever comportamento, não no curso do tempo de que gostaríamos que ela operasse. Mas, ainda assim, restaria a seguinte dúvida: quantas observações e amostras seriam suficientes para se obter uma boa aproximação?

Felizmente, em Estatística, o que não é conhecido pode ser estimado. Essa média que tanto nos interessa não foge à regra, graças ao teorema do limite central, ela também pode ser estimada a partir de uma única amostra de tamanho suficientemente grande e escolhida de maneira aleatória. A média de qualquer amostra que atenda a esses dois critérios terá baixa probabilidade de se desviar consideravelmente da média da população alvo[10]. Aliás, essa

[10] Não apenas a média; o desvio padrão de uma amostra representativa de uma determinada população também terá baixa probabilidade de se desviar consideravelmente do desvio padrão dessa mesma população.

é a ideia básica por trás do próprio teorema do limite central, que, além de tornar isso possível, ainda nos fornece o grau de confiança associado à média amostral. Lembre-se de que o teorema do limite central nos assegura que as médias amostrais estejam distribuídas de forma aproximadamente normal em torno da média da população (Figura 2). Ou seja, ele nos permite conhecer, a priori, a distribuição de probabilidade das médias amostrais em torno da média populacional: aproximadamente 68% de todas as médias amostrais estarão dentro do limite de 1 erro padrão da média populacional; cerca de 95%, dentro de 2 erros padrão; e 99%, dentro de 3 erros padrão (Figura 2). O erro padrão, que mede a variabilidade entre as médias amostrais, fornece a distância média do conjunto de todas as médias em relação à média populacional[11]. E é justamente essa previsibilidade que vai tornar possível a estimativa da média populacional a partir de uma única amostra de tamanho suficientemente grande e escolhida de maneira adequada.

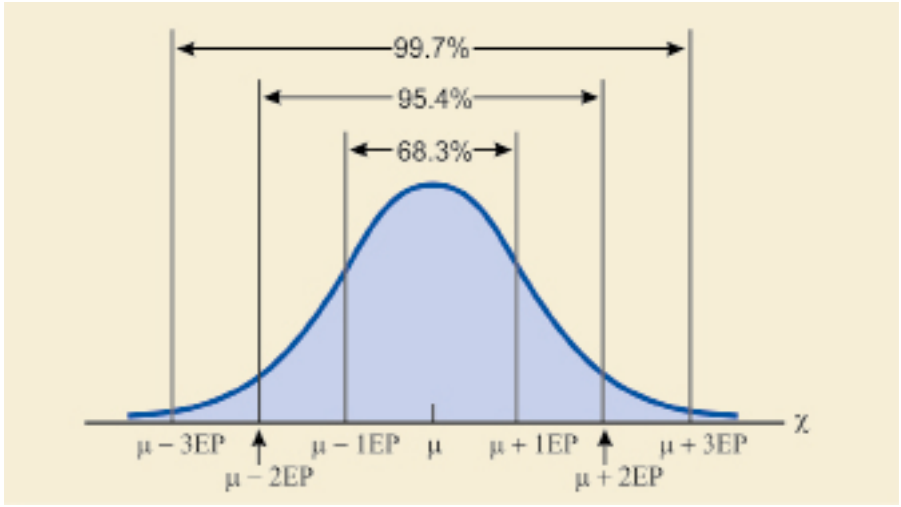


Figura 2. Curva de distribuição normal das médias amostrais
Fonte: ESCOLAEDIT (2018), modificada pelo autor.

[11] O leitor deve ter notado que o erro padrão têm o mesmo conceito básico de um desvio padrão: “ambos representam uma distância típica da média... Os valores da população original desviam-se uns dos outros graças a um fenômeno natural (as pessoas têm diferentes alturas, pesos, etc.), portanto temos o nome desvio padrão para medir sua variabilidade. As médias amostrais variam por causa do erro que ocorre por não sermos capazes de realizar um censo e temos que coletar amostras, portanto, temos o nome erro padrão para medir a variabilidade das médias amostrais”, Rumsey (2016, p. 159).

Apoiados nessa única amostra podemos então fazer a estimativa da média populacional e do erro padrão e, a partir desse último, se chegar à margem de erro, que nos daria uma ideia do quão essa média poderia variar, para mais ou para menos, se o processo de amostragem fosse repetido muitas outras vezes. Em geral, a margem de erro admitida é aquela em que uma média amostral esteja situada dentro do limite de 2 erros padrão para mais ou para menos da média populacional, o que nos daria uma ideia da acurácia de nossa média amostral[12]. Isso significa que a média populacional estaria dentro do intervalo estimado em 95% das vezes, se o processo de coleta de amostra fosse repetido muitas e muitas vezes. Ou seja, com base em uma única amostra, toma-se a decisão quanto à estimativa da média populacional, levando-se em consideração 95% de todos os resultados possíveis de ocorrer se novas amostras fossem coletadas. Mas isso também significa que, se o processo de amostragem fosse assim repetido, em 5% das vezes, e, por puro acaso, poder-se-ia selecionar amostras concentradas em elementos, e consequentemente médias, com valores atípicos muito baixos ou muito altos, aqueles situados nas caudas da curva de distribuição normal, que não representariam a população. Portanto, nessa situação e por puro acaso, a média populacional não seria captada pelo intervalo. Dissemos por puro acaso, porque, no mundo real, esses elementos com valores atípicos, que representam a menor parcela da população, geralmente não se encontram agrupados, assim como aparecem na ilustração da curva de distribuição normal, mas dispersos por toda a população. Portanto, na prática, em geral, eles também tendem a aparecer de forma dispersa ou espalhada em várias amostras de tamanho suficientemente grande e selecionadas de maneira aleatória, não tendo, por conseguinte, peso suficiente para inflar ou distorcer a média amostral. Por essa razão, a probabilidade de que uma amostra grande, selecionada de maneira aleatória, seja composta apenas por esses elementos atípicos é muito baixa, ou em torno de 5%. Quanto maior o nível de confiança desejado no valor da estimativa da média populacional, maiores serão a amplitude do intervalo de confiança da média e a margem de erro, uma vez que a distância até a média populacional, medida em erros padrão de ambos os lados, também aumentaria.

[12] Ao se adicionar e subtrair, de uma média amostral, um determinado número de erros padrão (em geral, 2 ou 3, o que também se conhece por margem de erro), obre-se o intervalo de confiança.

O brilhante físico inglês, Isaac Newton, foi o primeiro a utilizar a média como uma representação única de valores dispersos, embora ainda não tivesse conhecimento do teorema do limite central, uma vez que o mesmo ainda não havia sido plenamente desenvolvido. Nas palavras de Leonardo Mlodinow[13], autor do livro "O Andar do Bêbado", best-seller e livro notável do New York Times, pode-se encontrar, na página 136, o seguinte registro:

> "Uma maneira de gerar um número único a partir de uma série de medições discordantes é calcular a média. O jovem Isaac Newton parece ter sido o primeiro a empregá-la para esse propósito em suas investigações ópticas. Porém, como em tantas outras coisas, ele era uma exceção. A maioria dos cientistas nos tempos de Newton, e, no século seguinte, não calculava a média. Em vez disso, escolhia dentre suas medições um único "número áureo" – considerado, essencialmente por palpite, ser o mais confiável dos resultados obtidos".

Uma das coisas que o teorema do limite central também nos ensina é que uma das vantagens de se utilizar médias, em lugar do valor total ou de um único valor, para se fazer estimativas, decorre do fato de a variabilidade das médias amostrais diminuir à medida que o tamanho da amostra aumenta. E, ainda mais, que as médias amostrais sempre apresentarão menor variabilidade que os valores individuais dos quais foram gerados. É justamente por essas razões que o erro padrão da média, que mede a variabilidade entre as médias amostrais, é sempre menor que o desvio padrão, que, por sua vez, mede a variabilidade entre os elementos de uma amostra.

Uma aplicação prática do Teorema do Limite Central

Neste item, faremos uso do teorema do limite central para examinar a consistência entre os dados do Censo Agropecuário e os dados da Produção Agrícola Municipais (PAM), ambos do Instituto Brasileiro de Geografia e Estatística (IBGE). Para tanto, vamos utilizar, a título de exemplo, as médias das quantidades produzidas e da área colhida de mandioca na Bahia, no Pará e no Paraná, três dos quatro maiores produtores, calculadas com base nos

[13] Ver Mlodinow (2009, p. 136)

dados consolidados do Censo Agropecuário de 2006, que representarão as médias populacionais, e as médias das mesmas variáveis obtidas a partir dos dados da PAM, também de 2006, que farão o papel, por suposição, das médias amostrais dessa mesma população. Nesta primeira situação, parte-se do conhecimento sobre os parâmetros da população para verificar a consistência das estimativas calculadas a partir da amostra. Trata-se, portanto, de uma análise no sentido inverso do mais usual, quando, a partir de uma amostra, faz-se inferência sobre os parâmetros da população. A produtividade não será objeto desse tipo de análise, uma vez que, diferentemente da quantidade produzida e da área colhida, ela é uma estatística de aferição indireta. Em virtude disso, qualquer erro de medida, seja na produção ou na área, será refletido no cálculo da produtividade. A propósito, o IBGE também não faz levantamento de produtividade, seja no censo ou na PAM, mas a calcula como uma razão entre a quantidade produzida e a área colhida.

Uma vez que nem sempre o censo cobre, por problemas operacionais, exatamente 100% da população, também tomaremos, alternativamente e por suposição, os dados do Censo Agropecuário como uma grande, porque não dizer gigantesca, amostra. Nesse caso, a análise parte de duas amostras (o censo e a PAM) para se fazerem inferências sobre os parâmetros da população da qual elas foram supostamente extraídas.

Por natureza do próprio processo intrínseco na coleta, os dados do censo gozam de maior credibilidade do que os dados da PAM, porque se apoia em princípios estatístico dos mais fortes: a possibilidade de se conhecerem os verdadeiros, ou quase isso, parâmetros ao se trabalhar com toda ou quase toda a população, em vez de uma amostra, da qual só é possível se fazer inferências estatísticas sobre a população. Ainda assim, a acurácia da estatística ou a estimativa do parâmetro dependeria do tamanho da amostra e da maneira como ela foi escolhida. Por causa disso, os dados do Censo Agropecuário serão a nossa referência, seja dos dados populacionais, seja nos amostrais, embora, não devamos nos esquecer de que os dados do Censo Agropecuários são de natureza declarativa. Isso tira um pouco do poder e do brilho que se poderia esperar dos dados do Censo Agropecuário, uma vez que, por questões técnicas e operacionais, os mesmos não são fruto de aferições diretas, mas da declaração feita por cada informante censitário. Em virtude disso, reside no entrevistado, no caso o empresário rural, o produtor

ou o administrador do estabelecimento, o maior ônus atribuído ao grau de confiabilidade dos dados do censo. A premissa que se faz, e que também não poderia ser diferente, é que eles podem informar, com maior grau de precisão e acurácia (Figura 3), e como ninguém mais poderia, os resultados dos negócios com os quais trabalham e dependem, e que também não se furtariam de fazê-lo de forma fidedigna aos agentes censitários.

Figura 3. Diferença entre precisão e acurácia (ou exatidão)[14]
Fonte: BASEAEROFOTO (2018)

Os dados da PAM, por outro lado, apoiam-se em informações de produtores, de técnicos e dos próprios agentes do IBGE, com base na percepção, evidentemente incompleta e sem exatidão, que eles têm sobre a região onde atuam. Esse é mais um dos motivos para se verificar a consistência entre os dados da PAM e os dados do Censo Agropecuário de 2006, uma vez que ambos disponibilizam medidas sobre as mesmas coisas, no mesmo espaço geográfico e no mesmo período de tempo.

Alicerçados nas ideias apresentadas aqui, examinamos na Tabela 1 a consistência entre os dados de mandioca, no Estado da Bahia, do Censo Agropecuário e da PAM, tomando, inicialmente, o censo como a população e a PAM como uma amostra, tendo como unidade de investigação o muni-

[14] A precisão pode ser definida como o grau de proximidade entre os valores de uma variável obtidos pela repetição do processo de mensuração. Quanto menor for a diferença desses valores, melhor será a precisão. A acurácia ou exatidão, por sua vez, é o grau de proximidade da medida em relação ao verdadeiro valor da variável. Quanto mais próximo do valor real, melhor será a acurácia. Como diria Silver (2013, p. 53), ao exemplificar a diferença entre precisão e acurácia: “é como dizer que você tem uma boa mira porque seus tiros acertam mais ou menos o mesmo ponto – ainda que estejam longe de atingir o alvo”.

cípio. No Censo Agropecuário, de 2006, a quantidade média produzida de mandioca na Bahia foi de 2.098,74 toneladas, em uma área média colhida de 541,21 hectares. Assim, podemos esperar, fundamentados no teorema do limite central, que qualquer amostra grande o bastante e escolhida de maneira adequada, dessa mesma população de municípios produtores de mandioca na Bahia, apresentará médias próximas às médias do censo, por vezes, um pouco menor; por outras, um pouco maior. Mas, raramente, e somente por acaso, elas se desviariam excessivamente desses valores. O teorema do limite central nos ensina, alicerçado na regra empírica da distribuição normal, que os desvios aceitáveis são no máximo de três erros padrão. Mas a maioria dos pesquisadores, assim como também o fizemos, optam por um máximo de dois erros padrão, que nos dariam uma confiança de aproximadamente 95% de que o intervalo de estimativa da média conteria o verdadeiro valor da média populacional. Com três erros padrão, essa confiança subiria para aproximadamente 99%. Mas, em compensação, a margem de erro também subiria.

Com base nas médias das observações do censo e nos respectivos desvios padrão, podemos calcular o erro padrão esperado da média amostral para qualquer amostra adequadamente extraída da população de municípios produtores de mandioca. No caso da Bahia, o tamanho da amostra foi de 406, igual ao número total de municípios produtores de mandioca no levantamento da PAM, em 2006. Tomando o limite de dois erros padrão, a quantidade média de produção esperada, em qualquer amostra de 406 municípios produtores de mandioca na Bahia, deveria ficar entre 1.706,78 a 2.490,70 toneladas, enquanto a média da área colhida estaria situada entre 456,84 a 625,58 hectares, portanto, dentro de um intervalo que, uma vez sendo elas representativas da população, também contemplassem a média do censo (Tabela 1). As médias calculas a partir dos dados da PAM (10.822,65 toneladas, em quantidade produzida; e 848,95 hectares, em área colhida) não se encaixam nesse intervalo e estão demasiadamente afastadas das médias do censo (Tabela 1). Com base no teorema do limite central, podemos esperar que 95% de todas as médias amostrais se situariam dentro de dois erros padrão da média populacional. O afastamento da média de produção nos dados da PAM em relação à média do censo é da ordem de 44,51 erros padrão, enquanto o afastamento em relação à média da área colhida é de aproximadamente 7,29 erros padrão. Portanto,

em ambos os casos, a margem de erro supera, de forma inequívoca, qualquer limite estatisticamente aceitável. Em virtude disso, podemos inferir que as médias nos dados da PAM estão fora dos limites esperados para as estimativas das médias populacionais da produção e da área colhida de municípios produtores de mandioca na Bahia, em 2006, feitas a partir de uma amostra de tamanho suficientemente grande e aleatória. Isso também é um indicativo de que, muito provavelmente, os dados da PAM de 2006 não provêm de uma amostra representativa da produção e da área colhida de mandioca na Bahia.

Ainda com base nos limites esperados para a média de uma amostra a dois erros padrão da média censo, foi realizada uma estimativa de quais seriam os valores totais esperados da quantidade produzida e da área colhida, se a média da PAM fosse, de fato, proveniente de uma amostra representativa (Tabela 2). Para tanto, multiplicamos os valores esperados dos limites inferiores e superiores pelo número de elementos da população total (o número de municípios produtores de mandioca na Bahia, segundo o censo) para se ter uma ideia acerca dos totais esperados da produção e da área (esse que é o principal interesse do IBGE). Mais uma vez, como já era de se esperar, o total da produção e o total da área colhida, nos dados da PAM, estão demasiadamente fora dos limites esperados, sobretudo em relação à produção (Tabela 2).

Na Tabela 3, analisamos, de forma alternativa, a consistência entre as estatísticas de mandioca do Censo Agropecuário e da PAM, na Bahia, tomando ambos como dados amostrais de uma mesma população. Com base no teorema do limite central podemos esperar que, se colhêssemos 100 amostras, em 95 das vezes a média da nossa amostra estará dentro de dois erros padrão da média da população, para mais ou para menos. Portanto, com base nesse conhecimento, podemos esperar que os limites de variação das médias de duas amostras de uma mesma população se sobreponham. O intervalo de confiança da média da produção calculada a partir dos dados do censo seria de 1.706,28 a 2.491,20 toneladas por município. Fazendo uso desse mesmo procedimento, encontramos que o intervalo de confiança, a partir dos dados da PAM, foi de 8.872,05 a 12.773,25 toneladas por município. Quanto à área média colhida, os dados do censo nos forneceriam o intervalo de 456,73 a 625,69 hectares por município, enquanto os dados da PAM nos dariam 710,53 a 987,37 hectares por municípios. Percebemos, claramente,

em ambos os casos, que os intervalos de confiança não se sobrepõem. Percebemos, ainda, que os limites inferiores, nos dados da PAM, também em ambos os casos, são mais altos que os limites superiores calculados a partir dos dados do censo. Uma vez que a média populacional é um valor único, ela não pode estar contida, simultaneamente, em dois intervalos que não se sobrepõem. Portanto, podemos inferir que as médias do censo e da PAM não foram extraídas, ou não são representativas da média, de uma mesma população, ou seja, muito provavelmente ambas não poderiam, ao mesmo tempo, representar as médias de produção e da área colhida, em 2006, de mandioca no Estado da Bahia. Em virtude das razões já mencionadas que nortearam a escolha dos dados do censo como a nossa referência de dados mais acurados, podemos concluir que os dados da PAM não provêm de amostras representativas ou, dito de outra forma, que estas não reproduzem as mesmas características da população da qual foram extraídas. Podemos notar também que, as distorções nos dados da PAM são maiores nas estatísticas associadas às quantidades produzidas.

Os resultados do Estado do Pará estão mais consistentes que aqueles encontrados para o Estado da Bahia quando se toma o censo como a população e a PAM como uma amostra. Em que pese a média da quantidade produzida, obtida a partir dos dados da PAM, não se encaixa dentro do intervalo esperado para a média de uma amostra aleatória, a distância em erros padrão em relação à média do censo é de apenas 2,80, um pouco maior que o critério estabelecido neste trabalho, mas dentro da margem de 3 erros padrão, que é um limite aceitável em muitas situações (Tabela 4). A média da área colhida, por sua vez, situa-se aproximadamente, em valor absoluto, a 0,12 erro padrão da média-censo, portanto, abaixo do limite estabelecido, que foi de 2 erros padrão. Dessa forma, podemos inferir que a média da área colhida da PAM está dentro dos limites esperados para a média de uma amostra extraída da população de municípios produtores de mandioca no Pará, em 2006 (Tabela 4). A extrapolação da quantidade produzida e da área colhida, com base nas médias esperadas de uma amostra a dois-erros padrão da média do censo, pode ser encontrada na Tabela 5. Quando se parte da hipótese de que as médias de ambas as fontes de dados (censo e PAM) pertencem a amostras de uma mesma população, os resultados revelam que tal hipótese não pode ser negada (Tabela 6). Entretanto, no caso da média da quantidade produzida, a sobreposição dos intervalos de confiança a 95% é

apenas parcial: somente o limite inferior do intervalo da média-PAM cai dentro do intervalo da média-censo (Tabela 6).

Alicerçados no teorema do limite central, podemos ainda inferir que, se duas amostras são retiradas de uma mesma população, a diferença esperada entre suas médias seria próxima de zero. Mais uma vez, essa diferença poderia ser um pouco menor ou um pouco maior que zero, mas não se desviaria excessivamente disso, a não ser por puro acaso. Isso acontece porque o teorema do limite central nos assegura que, em amostras repetidas, a diferença entre duas médias terá distribuição aproximadamente normal, assim como acontece com as próprias médias que as geraram. Adotando o mesmo critério, que utilizamos para o caso das médias, poderíamos esperar que, em 95% das vezes, a diferença entre as médias das duas amostras estaria dentro do limite de dois erros padrão de zero, para mais ou para menos.

Uma vez que o resultado relacionado à quantidade produzida, no caso do Pará, apresenta uma certa ambiguidade, resolvemos, também, aplicar esse procedimento. Com base nessa metodologia, podemos observar que a distância de zero, medida em erros padrão, entre a diferença das médias PAM e censo foi de aproximadamente 1,91; portanto, dentro do limite aceitável (Tabela 7). Assim sendo, com base nesse procedimento, podemos também inferir que não há razão para negar a hipótese de que as médias de produção do censo e da PAM foram extraídas, ou são representativas, da média de uma mesma população.

O leitor atento deve ter percebido que a ambiguidade ainda persiste, ou seja: quando foram aplicados os procedimentos tomando os dados do censo como a população e os dados da PAM como uma amostra dessa mesma população, constatou-se, com 95% de confiança, que tal suposição não se sustentava, ou que a média da produção PAM estava fora dos limites esperados para a estimativa da média populacional feita a partir de uma amostra representativa e extraída da população de municípios produtores de mandioca no Pará, em 2006. Na tentativa de elucidar, definitivamente, essa contradição, os dois primeiros procedimentos foram repetidos, tomando-se os dados de ambas as fontes em logaritmos naturais. Esse processo de transformação dos dados teve por objetivo estabilizar a variância, a provável origem da causa da ambiguidade nos resultados. É como um paciente que precisa estabilizar suas

funções vitais antes de um procedimento cirúrgico para que a intervenção tenha maiores chances de sucesso. Ou seja, assim como a possibilidade de sucesso de um procedimento cirúrgico depende das condições iniciais em que se encontra o paciente, a possibilidade de sucesso na aplicação do teorema do limite central também depende do "comportamento" da variância. Esse mesmo processo não foi utilizado de forma generalizada para as demais séries por impossibilidade matemática: a presença do valor zero[15]. Apoiados nesse processo adicional de transformação dos dados e repetindo os dois primeiros procedimentos, constatamos que a ambiguidade deixa de existir, ou seja: agora é possível inferir que a média de produção nos dados da PAM não se encaixa nos limites esperados para a média de uma amostra representativa, em 2006, da população de municípios produtores de mandioca no Estado do Pará (Tabela 8), e, de forma análoga, que as médias do censo e da PAM, muito provavelmente, não foram extraídas, ou não são representativas da média, da mesma população (Tabela 9).

No caso do Estado do Paraná, em virtude do número de municípios no levantamento da PAM superar o número de municípios no levantamento do censo, a inconsistência matemática gerada nos impede que se faça a análise de consistência entre os dados dessas duas fontes, tomando o censo como a população e a PAM como uma amostra. Isso acontece por uma simples razão: um subconjunto não pode ser maior que o conjunto, assim como uma amostra não pode ser maior que a população. Dessa forma, a análise para o Paraná é feita apenas admitindo a segunda suposição: ambas as fontes contêm dados amostrais de uma mesma população. Assim como aconteceu com os dados da Bahia, notamos claramente na Tabela 10 que os intervalos de confiança da média do censo e da PAM, no Estado do Paraná, não se sobrepõem, seja em relação à produção ou à área colhida: os limites inferiores, nos dados da PAM, também em ambos os casos, são mais altos que os limites superiores nos dados do censo. Mais uma vez, isso é uma evidência de que as médias do censo e as da PAM não foram extraídas ou não são representativas da média de uma mesma população, ou seja, muito provavelmente, ambas não

[15] Por definição, sendo **a** e **b** números reais positivos, o logaritmo de **b** na base **a** é o expoente a que **a** deve ser elevado de modo que a potência obtida de base **a** seja igual a **b**. A expressão matemática correspondente seria: $\log_a b = x \Leftrightarrow a^x = b$; com $a > 0$, $a \neq 1$ e $b > 0$. Ou seja, não é possível elevar um número positivo **a** maior do que 1 a qualquer potência e obter 0.

poderiam, ao mesmo tempo, representar as médias de produção e da área colhida, em 2006, de mandioca no Estado.

De forma geral, notamos que os dados da PAM relacionados à lavoura de mandioca estão superestimados. Em geral, as discrepâncias são maiores na quantidade produzida que na área colhida, o que, por conseguinte, termina por superestimar também as produtividades da lavoura. Talvez isso decorra da maneira como os dados da PAM são levantados. Os técnicos, produtores e demais informantes que apoiam o IBGE nessa árdua e difícil tarefa[16], provavelmente são representantes de segmentos mais organizados do setor produtivo; aqueles com maior acesso à tecnologia, talvez, até mesmo por conta disso, com menor conhecimento sobre os segmentos menos dinâmicos, que, de forma geral, ainda representam uma grande parcela dos agricultores brasileiros, especialmente em relação à mandioca. Se esse realmente for o caso, teríamos estimativas feitas com base no conhecimento parcial que os informantes têm de uma parcela da produção que acontece em um segmento mais organizado e tecnificado, que não representaria a população de produtores em determinadas lavouras. Ou seja, a rigor, em situações como essas, os dados também não se enquadrariam no conceito estatístico de amostragem probabilística. Portanto, também a rigor, não deveriam ser feitas inferências estatísticas sobre a população a partir desses mesmos dados. Os exemplos apresentados aqui deixam expostos os riscos de se fazer inferências sobre a população a partir de uma amostra não aleatória: a perda da acurácia.

Todo sistema de medição precisa passar uma imagem clara da realidade. Nada é observado fora do filtro de algum tipo de sistema de medição, seja ele abstrato ou concreto. Aliás, essa também é uma área de estudo da Estatística conhecida como Análise da Capacidade do Sistema de Medição, ou simplesmente MSA. Em 2017, uma outra fonte de discordância entre os dados da PAM e do censo está, provavelmente, relacionada aos distintos períodos de referência dos dados, embora na divulgação dos resultados, as estatísticas se reportem ao mesmo ano. Enquanto no Censo Agropecuário de 2006, o período de referência, que também nos parece mais adequado, foi

[16] Ainda mais difícil é executar essa tarefa mediante uma conjuntura desfavorável de redução de quadro de pessoal e do enfraquecimento da rede de instituições de apoio ao levantamento de dados em todo o País, em especial as empresas estaduais de extensão rural.

de 1º de janeiro a 31 de dezembro daquele ano; no censo de 2017, optou-se por 1º de outubro de 2016 a 30 de setembro de 2017. Ou seja, o período de referência atual do censo não coincide com o estabelecido pela PAM, que, por sua vez, utiliza o ano civil, tal como se fez no censo de 2006.

Apesar dessa mudança, o Censo Agropecuário ainda constitui uma boa referência para os dados da PAM, mas, infelizmente, o processo só acontece a cada dez anos. Então, à medida que o tempo passa, ele vai se tornando mais fraco como uma referência ou linha de base, especialmente para as lavouras temporárias. É como o farol de um carro visto pelo retrovisor de um veículo que dirigimos em uma estrada: à medida que nos afastemos dele (do carro que vem logo atrás) a luz projetada no retrovisor vai ficando cada vez mais fraca e, em um determinado ponto, deixa de incomodar e, depois de mais um tempo, até desaparece. Assim, também, acontece com o censo, enquanto base de referência para as estimativas da PAM. No início, os técnicos do IBGE podem e fazem uso dos dados do censo para calibrar as opiniões dos informantes da PAM. Mas, à medida que o tempo passa, tudo pode mudar, então os mesmos informantes se sentem menos “incomodados” para opinar livremente a respeito das mesmas questões. Portanto, a redução na periodicidade do censo poderia ser um bom caminho para melhor nortear as estimativas da PAM. Mas isso teria um alto custo e dependeria de decisão governamental. Outra opção seria mudar a metodologia de estimativa da PAM. Isso seria possível com o uso de tecnologia, tal como a utilização de imagem de satélite no levantamento da área plantada. Conhecendo-se a área plantada, uma grande parte do problema terá sido resolvida e será mais seguro fazer estimativas de produção a partir daí. Ou, alternativamente, com o emprego de uma combinação de métodos envolvendo a utilização de técnicas adequadas de amostragem probabilística acerca da área colhida e da quantidade produzida. Isso também vai custar caro, mas, com certeza, os benefícios vão compensar os custos. Informação com precisão e acurácia é quase tudo no planejamento e na tomada de decisão.

Tabela 1. Análise de consistência entre os dados do Censo Agropecuário e da PAM, tomando o censo como a população e a PAM como uma amostra. Lavoura: mandioca. Unidade da Federação: Bahia. Ano, 2006.

Variável	Média Censo	Média PAM	Diferença (PAM - Censo)	Desvio Padrão (Censo)	Erro Padrão Esperado da Média da Amostra[a]	Margem de Erro Esperada da Média da Amostra	Limites Esperados para a Média da Amostra a Dois Erros padrão da Média da População[a]	Distância Média-PAM da Média-Censo (em número de erros padrão)	Viés (PAM)	Decisão
Produção (em toneladas)	2.098,74	10.822,65	8.723,91	3.948,98	195,98	391,96	1.706,78 a 2.490,70	44,51	↑	A média da produção da PAM está fora dos limites esperados, a dois erros padrão da média da população, da média de uma amostra extraída de municípios produtores de mandioca na Bahia, em 2006.
Área colhida (em hectares)	541,21	848,95	307,74	850,04	42,19	84,37	456,84 a 625,58	7,29	↑	A média da área colhida da PAM está fora dos limites esperados, a dois erros padrão da média da população, da média de uma amostra extraída da população de municípios produtores de mandioca na Bahia, em 2006.

[a] Valor esperado para qualquer amostra aleatória de tamanho (n) 406, que é igual ao número de elementos ou municípios no levantamento da PAM - nossa amostra, por suposição.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 2. Totais encontrados e totais esperados nos dados da PAM, com base nos limites do intervalo de confiança esperado da média amostral. Lavoura: mandioca. Unidade da Federação: Bahia. Ano, 2006.

Variável	Total PAM	Total Censo	Número de Municípios (Censo)[a]	Limites Esperados para a Média da Amostra a Dois Erro - Padrão da Média da População[a]	Limite Inferior Esperado para o Total (PAM)	Limite Superior Esperado para o Total (PAM)
Produção (em toneladas)	4.393.997,00	852.090,00	406	1706,78 a 2490,70	692.752,00	1.011.224,00
Área colhida (em hectares)	344.672,00	219.732,00	406	456,84 a 625,58	185.477,00	253.985,00

[a] O número de municípios no levantamento do censo coincide com o número de municípios no levantamento da PAM, que foi igual a 406.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 3. Análise de consistência entre as estatísticas do Censo Agropecuário e as da PAM, tomando ambos como dados amostrais de uma mesma população. Lavoura: mandioca. Unidade da Federação: Bahia. Ano, 2006.

Variável	Média Censo	Média PAM	Diferença (PAM – Censo)	Desvio Padrão (Censo) [a]	Desvio Padrão (PAM) [a]	Erro Padrão (Censo)	Erro Padrão (PAM)	Margem de Erro da Média (Censo)	Margem de Erro da Média (PAM)	Intervalo de Confiança da Média a 95% (Censo)	Intervalo de Confiança da Média a 95% (PAM)	Decisão
Produção (em toneladas)	2.098,74	10.822,65	8723,91	3.953,85	19.651,85	196,23	975,30	392,46	1.950,60	1.706,28 a 2.491,20	8.872,05 a 12.773,25	Não há sobreposição dos intervalos de confiança das **médias** do censo e da PAM.
Área colhida (em hectares)	541,21	848,95	307,74	851,09	1.394,58	42,24	69,21	84,48	138,42	456,73 a 625,69	710,53 a 987,37	Não há sobreposição dos intervalos de confiança das **médias** do censo e da PAM.

[a] Número de municípios no levantamento da PAM igual ao do censo: 406.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 4. Análise de consistência entre os dados do Censo Agropecuário e da PAM, tomando o censo como a população e a PAM como uma amostra. Lavoura: mandioca. Unidade da Federação: Pará. Ano, 2006.

Variável	Média Censo	Média PAM	Diferença (PAM - Censo)	Desvio Padrão (Censo)	Erro Padrão Esperado da Média da Amostra[a]	Margem de Erro Esperada da Média da Amostra	Limites Esperados para a Média da Amostra a Dois Erros padrão da Média da População[a]	Distância entre a Média-PAM e a Média-Censo (em números de erros padrão)	Viés (PAM)	Decisão
Produção (em toneladas)	22.451,93	37.898,70	15.446,77	63.904,53	5.520,51	11.041,02	11.410,91 a 33.492,95	2,80	↑	A média da produção PAM está fora dos limites esperados, a dois erros padrão da média da população, da **média de uma amostra** extraída da população de municípios produtores de mandioca no Pará, em 2006.
Área colhida (em hectares)	2.408,52	2.343,85	-64,67	6.130,11	529,56	1.059,12	1.349,40 a 3.467,64	-0,12	Ausente	A média da área colhida da PAM está dentro dos limites esperados, a dois erros padrão da média da população, da **média de uma amostra** extraída da população de municípios produtores de mandioca no Pará, em 2006.

[a] Valor esperado para qualquer amostra aleatória de tamanho (n) 134, que é igual ao número de elementos ou municípios no levantamento da PAM - nossa amostra, por suposição.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 5. Totais encontrados e totais esperados nos dados da PAM, com base nos limites do intervalo de confiança esperado da média amostral. Lavoura: mandioca. Unidade da Federação: Pará. Ano, 2006.

Variável	Total PAM	Total Censo	Número de Municípios (Censo)[a]	Limites Esperados para a Média da Amostra a Dois Erro - Padrão da Média da População[a]	Limite Inferior Esperado para o Total (PAM)	Limite Superior Esperado para o Total (PAM)
Produção (em toneladas)	5.078.426,00	3.075.915,00	137	11.370,39 a 33.533,47	1.557.743,00	4.594.085,00
Área colhida (em hectares)	314.076,00	329.967,00	137	1.345,51 a 3.471,52	184.335,00	475.598,00

Fonte: Cálculo do autor, com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 6. Análise de consistência entre as estatísticas do Censo Agropecuário e as da PAM, tomando ambos como dados amostrais de uma mesma população. Lavoura: mandioca. Unidade da Federação: Pará. Ano, 2006.

Variável	Média Censo	Média PAM	Diferença (PAM – Censo)	Desvio Padrão (Censo) [a]	Desvio Padrão (PAM) [a]	Erro Padrão (Censo)	Erro Padrão (PAM)	Margem de Erro da Média (Censo)	Margem de Erro da Média (PAM)	Intervalo de Confiança da Média a 95% (Censo)	Intervalo de Confiança da Média a 95% (PAM)	Decisão
Produção (em toneladas)	22.451,93	37.898,70	15.446,77	64.139,04	69.157,68	5.479,77	5.974,31	10.959,54	11.948,62	11.492,39 a 33.411,47	25.950,08 a 49.847,32	Há sobreposição parcial dos intervalos de confiança entre as médias do censo e as da PAM.
Área colhida (em hectares)	2.408,52	2.343,85	-64,67	6.152,60	3.780,61	525,65	326,59	1.051,30	653,18	1.357,22 a 3.459,82	1.690,67 a 2.997,03	Há sobreposição dos intervalos de confiança entre as **médias** do censo e as da PAM.

[a] Número de municípios no censo, igual a 137. b Número de municípios no levantamento da PAM, igual a 134.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 7. Análise de consistência entre as médias da produção do Censo Agropecuário e da PAM, tomando ambos como dados amostrais de uma mesma população. Lavoura: mandioca. Unidade da Federação: Pará. Ano, 2006.

Variável	Média--Censo	Média--PAM	Diferença (PAM – Censo)	Desvio Padrão (Censo)	Desvio Padrão (PAM)	Número de Municípios (Censo)	Número de Municípios (PAM)	Erro Padrão para Diferença Entre Médias	Distância de Zero da Diferença entre a Média--PAM e a Média-Censo (em número de erros padrão)	Decisão
Produção (em toneladas)	22.451,93	37.898,70	15.446,77	64.139,0400	69.157,6821	137	134	8.106,8047	1,91	A diferença entre as **médias** de produção do censo e as da PAM estão dentro do limite esperado.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 8. Análise de consistência entre as estatísticas de produção do Censo Agropecuário e as da PAM, tomando o censo como a população e a PAM como uma amostra. Dados expressos em logaritmos naturais. Lavoura: mandioca. Unidade da Federação: Pará. Ano, 2006.

Variável	Média-Censo	Média-PAM	Diferença (PAM - Censo)	Desvio Padrão (Censo)	Erro Padrão Esperado da Média da Amostra[a]	Margem de Erro Esperada da Média da Amostra	Limites Esperados para a Média da Amostra a Dois Erros padrão da Média da População[a]	Distância Média-PAM da Média-Censo (em número de erros padrão)	Viés (PAM)	Decisão
Produção (em toneladas)	7,927439	9,586553	1,659114	2,380150	0,205614	0,411228	7,516211 a 8,338667	8,069071	↑	A média da produção PAM está fora dos limites esperados, a dois erros padrão da média da população, da média de uma amostra extraída da população de municípios produtores de mandioca no Pará, em 2006.

[a] Valor esperado para qualquer amostra aleatória de tamanho (n) 134, que igual ao número de elementos ou municípios no levantamento da PAM - nossa amostra, por suposição.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 9. Análise de consistência entre as estatísticas de produção do Censo Agropecuário e as da PAM, tomando ambas como dados amostrais de uma mesma população. Dados expressos em logaritmos naturais. Lavoura: mandioca. Unidade da Federação: Pará. Ano, 2006.

Variável	Média-Censo	Média-PAM	Diferença (PAM – Censo)	Desvio Padrão (Censo) [a]	Desvio Padrão (PAM) [b]	Erro Padrão (Censo)	Erro Padrão (PAM)	Margem de Erro da Média (Censo)	Margem de Erro da Média (PAM)	Intervalo de Confiança da Média a 95% (Censo)	Intervalo de Confiança da Média a 95% (PAM)	Decisão
Produção (em toneladas)	7,927439	9,586553	1,659114	2,388885	1,476774	0,204096	0,127574	0,408192	0,255148	7,519247 a 8,335631	9,331405 a 9,841701	Não há sobreposição dos intervalos de confiança das **médias** do censo e as da PAM.

[a] Número de municípios no censo, igual a 137. b Número de municípios no levantamento da PAM, igual a 134.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Tabela 10. Análise de consistência entre as estatísticas do Censo Agropecuário e as da PAM, tomando ambos como dados amostrais de uma mesma população. Lavoura: mandioca. Unidade da Federação: Paraná. Ano, 2006.

Variável	Média-Censo	Média-PAM	Diferença (PAM – Censo)	Desvio Padrão (Censo) [a]	Desvio Padrão (PAM) [b]	Erro Padrão (Censo)	Erro Padrão (PAM)	Margem de Erro da Média (Censo)	Margem de Erro da Média (PAM)	Intervalo de Confiança da Média a 95% (Censo)	Intervalo de Confiança da Média a 95% (PAM)	Decisão
Produção (em toneladas)	3.406,42	9.722,44	6.316,02	7.145,45	19.526,30	373,50	982,48	747,00	1.964,96	2.659,42 a 4.153,42	7.757,48 a 11.687,40	Não há sobreposição dos intervalos de confiança entre as médias do censo e as da PAM.
Área colhida (em hectares)	255,37	437,85	182,48	475,68	826,02	24,86	41,56	49,72	83,12	205,65 a 305,09	354,73 a 520,97	Não há sobreposição dos intervalos de confiança entre as médias do censo e as da PAM.

[a] Número de municípios no censo, igual a 366. b Número de municípios no levantamento da PAM, igual a 395.

Fonte: Cálculo do autor com base nos dados básicos do IBGE (2018a e 2018b).

Referências

BASEAEROFOTO. **Qual a diferença entre precisão e acurácia?**. Disponível em http://www.baseaerofoto.com.br/faq/. Acesso em: 02 out. 2018.

DEMING, W. E. **Saia da Crise**: as 14 lições definitivas para controle de qualidade. São Paulo: Futura, 2003.

ELLENBERG, J. **O poder do pensamento matemático**: a ciência de como não estar errado. Rio de Janeiro: Zahar, 2015.

ESCOLAEDIT. **Afinal, de onde vem os 3,4 PPMS do Six Sigma?** Disponível em <:https://www.escolaedti.com.br/de-onde-vem-os-34-ppms/>. Acesso em: 20 set. 2018.

IBGE. **Censo Agropecuário de 2006**. Disponível em:http:// https://sidra.ibge.gov.br/tabela/825. Acesso em: 13 set. 2018a.

IBGE. **Produção Agrícola Municipal**. Disponível em: https://sidra.ibge.gov.br/tabela/1612. Acesso em: 14 set. 2018b.

MLODINOW, L. **O andar do bêbado**: como o acaso determina nossas vidas. Rio de Janeiro: Jorge Zahar, 2009.

MONTGOMERY, D. C. **Introdução ao controle estatístico da qualidade**. 7. ed: Rio de Janeiro: LTC, 2017.

RODRIGUES, M.V. Entendendo, aprendendo, desenvolvendo qualidade padrão seis **Sigma**. Rio de Janeiro: QualityMark, 2006.

RUMSEY, D. **Estatística para leigos**: uma maneira fácil e divertida de entender estatística! Rio de Janeiro: Alta Books, 2016.

SALSBURG, D. **Uma senhora toma chá**...: como a estatística revolucionou a ciências no século XX. Rio de Janeiro: Zahar, 2009.

SILVER, N. **O sinal e o ruído**: por que tantas previsões falham e outras não. Rio de Janeiro: Intrínseca, 2013.